DOCUMENT RESUME

ED 314 460 TM 014 326

AUTHOR Pike, Gary; Banta, Trudy
TITLE Using Construct Validity To Evaluate Assessment Instruments: A Comparison of the ACT-COMP Exam and the ETS Academic Profile.
INSTITUTION Tennessee Univ., Knoxville. Center for Assessment Research and Development.
REPORT NO RR-89-06
PUB DATE Mar 89
NOTE 50p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Ability; *Achievement Tests; *College Seniors; Comparative Testing; *Construct Validity; Educational Assessment; *Evaluation Methods; Higher Education; Outcomes of Education; Scores; Standardized Tests; Standards; Test Interpretation; Test Results; Test Validity
IDENTIFIERS *Academic Profile (ETS); *College Outcome Measures Project; Tennessee Higher Education Commission; Test Appropriateness; University of Tennessee Knoxville

ABSTRACT

The purpose of this paper is (1) to discuss a set of standards that can be used to evaluate potential assessment instruments; and (2) to use these standards to evaluate the American College Testing Program's College Outcomes Measures Program (ACT-COMP) and the Educational Testing Service (ETS) Academic Profile. Using the work of S. Messick (1975, 1987, 1988) on construct validity, researchers examined the substantive, structural, and external components of use of these tests by the Tennessee Higher Education Commission (THEC) and the University of Tennessee (Knoxville). The COMP was administered to 1,828 seniors, and the Academic Profile was administered to 1,173 seniors. Also, 35 seniors agreed to take both examinations. The Academic Profile was superior in its ability to differentiate accurately among students and programs. Both tests appeared to measure a single underlying construct, and analysis suggests that this construct is academic ability, not program quality. That the THEC defines effective general education in terms of test scores is particularly troubling in light of these findings. It is contended that the THEC guidelines may: (1) unjustifiably limit the substance of general education to a narrow range of learning outcomes; (2) award funds on the basis of differences that are well within the error of measurement for the examinations; and, most importantly (3) evaluate programs on the basis of the students they attract rather than quality of the education those students receive. Eleven tables of comparisons and one figure illustrating Messick's concept of validity are provided. (SLD)

CENTER FOR ASSESSMENT RESEARCH AND DEVELOPMENT

-- RESEARCH REPORT --

RR 89 - 06

Using Construct Validity to Evaluate Assessment Instruments: A Comparison of the ACT-COMP Exam and the ETS Academic Profile

by

Gary Pike & Trudy Banta

Center for Assessment Research and Development
The University of Tennessee, Knoxville
1819 Andy Holt Avenue
Knoxville, Tennessee 37996-4350
(615) 974-2350

USING CONSTRUCT VALIDITY TO EVALUATE ASSESSMENT INSTRUMENTS: A COMPARISON OF THE ACT-COMP EXAM AND THE ETS ACADEMIC PROFILE

Gary R. Pike
Associate Director

Trudy W. Banta
Research Professor & Director

Assessment Resource Center
2046 Terrace Avenue
University of Tennessee, Knoxville
2046 Terrace Avenue
Knoxville, TN 37996-3504
(615) 974-0883

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, March, 1989.

Abstract

Responding to criticisms of American higher education, state higher education commissions, along with the federal government and the regional accrediting associations, have taken actions to promote outcomes assessment and enhance educational quality. Because achievement tests are an integral part of outcomes measurement, it is not surprising that the growing interest in assessment is reflected in a dramatic increase in the number of available exams. Unfortunately, criteria for evaluating achievement tests as measures of program effectiveness are not clearly established. The purpose of this paper is to: (1) describe a set of standards that can be used to evaluate potential assessment instruments; and (2) to use these standards to evaluate the COMP exam and the Academic Profile.

Using Messick's work on construct validity, this research examined the substantive, structural, and external components of test use by the Tennessee Higher Education Commission (THEC) and the University of Tennessee, Knoxville (UTK). Results indicate that both tests are almost equal in their coverage of the UTK general education goals (30%). Regarding the structural component of test use, the Academic Profile is superior to the COMP exam in its ability to accurately differentiate among students/programs. Analysis of the structural component also reveals that both tests measure a single underlying construct, and analysis of the external component suggests that this construct is academic ability, not program quality. Analysis of these tests also suggests that the COMP exam is somewhat more sensitive to educational effects than the Academic Profile, but only after the effects of ability are removed.

The fact that the THEC defines effective general education in terms of test scores is particularly troubling in light of these findings. These results suggest that the THEC guidelines unjustifiably limit the substance of general education to a narrow range of learning outcomes, award funds on the basis of differences that are well within the error of measurement for the exams, and, most important, evaluate programs on the basis of the quality/ability of the students they attract, not the quality of the education those students receive.

USING CONSTRUCT VALIDITY TO EVALUATE ASSESSMENT INSTRUMENTS: A COMPARISON OF THE ACT-COMP EXAM AND THE ETS ACADEMIC PROFILE

Since 1983, six national advisory commissions have issued reports criticizing the quality of American higher education and suggesting that colleges and universities develop programs to assess student outcomes as part of an overall strategy to improve educational quality (Banta, 1988b). In response, state legislatures and higher education coordinating boards, along with the federal government and the regional accrediting associations, have taken actions intended to promote outcomes assessment (Ewell & Lisensky, 1988; National Governors' Association, 1988). Given this interest in assessment, it is not surprising that the American Council on Education's 1988 Campus Trends survey found that assessment activities were underway at two-thirds of the public institutions and 40% of the independent colleges surveyed (El-Khawas, 1988).

In gathering data about their education programs, colleges and universities rely on a variety of measurement techniques. While achievement tests represent only one part of an overall assessment effort that includes satisfaction surveys and performance appraisals, these tests play a major role in most assessment programs (Banta & Fisher, 1987; Harris, 1986).

Because achievement tests are an integral part of outcomes measurement, it is not surprising that the growing interest in assessment is reflected in a dramatic increase in commercially-available and locally-developed exams (Banta & Schneider, 1988; Pike, 1988). With increased test availability has come the problem of deciding which tests are most appropriate as assessment instruments. Unfortunately, criteria for evaluating achievement tests as measures of program effectiveness are not clearly established.

The purpose of this paper is twofold: (1) to describe a set of standards and a methodology that can be employed to evaluate the appropriateness of using achievement tests as assessment instruments; and (2) to use these standards to evaluate the ACT College Outcome Measures Program (COMP) examination and the ETS Academic Profile as measures of general education program effectiveness at the University of Tennessee, Knoxville (UTK). (It should be noted that the version of the Academic Profile used in this research was a pilot test and was replaced by the Academic Profile II in Fall 1988.) While this research focuses on standardized achievement tests, the procedures described in this paper can also be used to evaluate faculty-developed exams.

Standards for Evaluating Achievement Tests

The concept of validity provides a starting point for evaluating achievement tests as assessment instruments because validity is concerned with the accuracy and appropriateness of a basic component of assessment, the inferences and actions suggested by test scores (Cronbach, 1971; Messick, 1988b). As Millman (1988) explains, what is being validated is an interpretation, not a test. Thus, a test can be valid for one use or for one institution, but not for another use or another institution.

Validation research can be based on a variety of paradigms. For example, the most recent edition of the Standards for Educational and Psychological Testing identifies three "types" of validity: content, construct, and criterion-related (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). While distinctions between different types of validity can be useful, there is a growing sentiment that validity should be treated as a unitary concept with construct validity at its core (Messick, 1987).

Supporters of a unitary concept of validity do not discount the importance of content and criterion-related evidence. Instead, they argue that these types of evidence must be bolstered by evidence of construct validity. For example, judgments about content validity do not provide evidence about whether the test actually measures the domain its content seems to represent. Evidence of construct validity is required to determine if a test actually measures a domain (Messick, 1988a). Similarly, significant relationships between test scores and criterion variables do not provide evidence that the criterion variables are valid indicators of the construct they are assumed to represent. Again, evidence of construct validity is required.

Messick (1987) has identified two interconnected facets of validity. The first facet draws a distinction between evaluations of evidence and evaluations of consequences; the second draws a distinction between questions of score interpretation and questions of score use. When these facets are crossed, they produce the four-fold progressive matrix presented in Figure 1.

Insert Figure 1 about here

According to Messick (1987), construct validity provides the data used to make the judgments represented by each cell of the matrix. For the upper left-hand cell,

only construct validity can provide evidence about the appropriateness of score interpretations. The lower left-hand cell, the consequences of test interpretation, also requires evidence of construct validity. However, the focus of these evaluations is on the value implications of score interpretations (Messick, 1975). The upper right-hand cell represents questions about the evidential basis for test use. Here again, construct validation research provides the needed data, but this data must be specific to a particular test use. The final cell represents questions concerning the consequences of score use and draws on data used to answer questions in the first three cells. A test use is invalid in terms of its social consequences only if adverse consequences are the result of previously identified sources of invalidity (Messick, 1988a).

When colleges and universities undertake validity studies, the focus of these investigations is on test use, irrespective of whether the test is used to classify students or evaluate academic programs. Obviously, the first step in construct validation research involves delineating the construct(s) the test is intended to measure. Once a construct has been identified, a variety of approaches can be used to evaluate relationships between test scores and the construct. Loevinger (1957) has identified three components that guide the analysis of evidence concerning test use: (1) the extent to which the content of the test accurately represents the construct upon which actions are based (the substantive component); (2) the extent to which test structure accurately represents the structure of the construct (the structural component); and (3) the extent to which relationships between test scores and other variables are consistent with relationships implied by the construct (the external component).

Taken together, these components provide the information needed to determine if a particular test use avoids two threats to construct validity. The first, construct underrepresentation, occurs when a test does not adequately sample from the domain(s) of the construct, and the second threat, construct irrelevant variance, occurs when a test samples domains outside the construct, creating variance in test scores that is not related to the corstruct. Failure to provide support for the substantive, structural, or external components of construct validity is significant only if this lack of support is manifest in construct underrepresentation or construct irrelevant score variance.

An Empirical Example

The Context

The setting for this research is the University of Tennessee, Knoxville (UTK), the state's public research university. UTK has an enrollment of almost 20,000 undergraduate and 6000 graduate/professional students (University of Tennessee Office of Management Services, 1987). Undergraduates major in one of ten colleges: Agriculture, Architecture, Business, Communications, Education, Engineering, Human Ecology, Liberal Arts, Nursing, and Social Work. The general education program at UTK relies on a combination of distribution requirements (Centra, 1988), and each college defines the combination for its majors.

The campus-wide assessment program gathers information on educational outcomes by using achievement tests in general education, achievement tests in the major, and opinion surveys. Assessment findings also are used in annual budget hearings and periodic academic program reviews (Pike & Banta, 1987).

UTK is an active participant in the performance funding program administered by the Tennessee Higher Education Commission (THEC). As early as 1975 the THEC began discussing the possibility of basing a part of higher education funding on performance criteria, and in 1983 the commission established a set of performance funding guidelines which currently provide a financial supplement of up to 5% of an institution's budget for instruction based on the results of a series of evaluation activities (Levy, 1986; Pike & Banta, 1987). In 1986, the THEC voted to continue the performance funding program and established new criteria for judging institutional performance in the areas of program accreditation, student learning in the major, student learning in general education, alumni satisfaction, and corrective (improvement) measures (Banta, 1988a).

The performance funding standard on learning in general education is the most widely publicized THEC guideline and determines one-fifth of an institution's funding supplement. It is difficult to determine what constructs are being evaluated by the standard on general education because the THEC does not identify the domains of general education, nor does it define what is meant by program effectiveness. By default, the THEC procedures for awarding funds and the test used to measure educational outcomes may have become the general education construct on some public campuses in Tennessee.

In the area of general education, the commission awards funds on the basis of student performance on the ACT-COMP exam. Performance on the COMP exam is judged by institutional means (national percentile ranks) and measures of the value added by

general education from the freshman to senior years (Banta, 1988a). In addition, the performance funding standard on corrective measures requires that institutions use subscores on the COMP exam to implement program changes that will improve total scores on the exam.

While the THEC has not defined the constructs comprising general education, UTK has a specific statement describing general education content areas and student development goals. In 1981, the Coordinating Committee on General Education issued a report identifying three general education components: basic skills, areas of knowledge and patterns of inquiry, and attitudes and perceptions (Humphreys, 1984). Each of these components is subdivided into specific content areas (see Table 3).

The most recent revision of the performance funding guidelines provides extra credit for developing and/or pilot testing new assessment instruments. During the 1987-88 academic year, UTK participated in pilot testing the Academic Profile. What follows is a comparison of the COMP exam and the Academic Profile. Because problems with the validity of gain scores have been reported elsewhere (Banta, Lambert, Pike, Schmidhammer, & Schneider, 1987), this research focuses on the validity of raw score use.

The Instruments

The COMP exam and the Academic Profile are among the most widely used measures of general education knowledge and skills. Because it is difficult, if not impossible, to identify a core of knowledge common to general education programs at most colleges and universities, both tests minimize the need to recall specific facts. However, the staff at both ACT and ETS contend that familiarity with content does improve test performance.

The COMP exam is available in two forms: the Objective Test (containing multiple-choice questions) and the longer Composite Examination (consisting of multiple-choice items and exercises requiring students to write essays and record speeches) (Forrest & Steele, 1982). Most institutions, including UTK, use the Objective Test because it is easier to administer and score (Banta et al., 1987).

The Objective Test takes approximately 2 1/2 hours to administer and contains 60 questions, each with two correct answers. The questions are divided among 15 separately timed activities drawing on material (stimuli) from television programs, radio broadcasts, and print media. Students taking the COMP exam are instructed that there is a penalty for guessing (i.e., incorrect answers will be subtracted from their scores), but that leaving a question blank will not be counted against

them. The combination of two correct answers for each question, the guessing penalty, and no penalty for not answering a question means that the score range for each of the 60 items is from -2 to 2. A score of -2 represents two incorrect answers, while a score of -1 represents one incorrect answer and one left blank. A score of 0 can represent either both answers left blank or one correct and one incorrect answer. A score of 1 represents one correct answer and a blank, and a score of 2 represents two correct answers. The range for item scores is rescaled to 0 to 4 points, making the maximum possible score on the Objective Test 240 points and a chance score 120 points.

In addition to a total score, the COMP exam provides three content subscores (Functioning within Social Institutions [FSI], Using Science and Technology [US], and Using the Arts [UA]) and three process subscores (Communicating [COM], Solving Problems [SP], and Clarifying Values [CV]). Alpha reliability is estimated to be .84 for the total score and to range from .63 to .68 for the subscores (Forrest & Steele, 1982). It is difficult to determine precisely what constructs these subscales are designed to measure because the technical manual for the COMP exam provides only one-paragraph descriptions of the subscales.

The Academic Profile also is available in two forms. The three-hour version of the test provides scores for individuals, while the one-hour form uses a matrix-sampling procedure designed to provide institutional score reports. Like the COMP exam, the Academic Profile uses multiple stimuli. However, all of the stimuli in the Academic Profile are presented in written passages.

The three-hour version of the Academic Profile was used at UTK and consists of 144 questions designed to measure four skills (Reading [READ], Writing [WRITE], Critical Thinking [CT], and Mathematics [MATH]) across three content areas (Humanities [HUM], Social Sciences [SS], and Natural Sciences [NS]). Each question is scored as either right or wrong, without a guessing penalty, making the maximum possible score on the exam 144 points. According to ETS, estimates of alpha reliability are .94 for the total score and range from .78 to .85 for the seven subscores (Dick Burns, personal communication, January 9, 1989)[1]. As is the case with the COMP exam, ETS provides little information about the constructs the Academic Profile is designed to measure.

The Research Questions

In order to demonstrate the construct validity of the COMP exam and the Academic Profile when used as assessment instruments, analyses of the substantive, struc-

tural, and external components of construct validity must provide evidence that these exams are capable of identifying the strengths and weaknesses of UTK's general education program and must suggest specific actions that can be taken to improve program quality.

Evaluating the substantive component of construct validity involves determining whether the COMP exam and the Academic Profile accurately represent the domains of the general education program. Judgments about content representativeness were made by two groups of reviewers: faculty and students. Faculty reviewers consisted of seven faculty members drawn from five undergraduate colleges who reviewed the tests and then met with the authors to arrive at a consensus about test coverage. Students were asked to rate the content coverage of the exams after they had taken them.

Evaluating the structural component of construct validity involved comparing the structure of the tests to the structure of the constructs they are designed to measure. Three questions related to test structure formed the basis for evaluating the COMP exam and the Academic Profile: the appropriateness of the exams' norm-referenced scoring procedures, the reliability/generalizability of test scores, and the dimensionality of subscores.

Evaluating the external component of construct validity involved determining the sensitivity of the two tests to students' educational experiences. Obviously, a test that is used to award funds and suggest specific program changes should reflect the effects of a general education curriculum (Pike, 1988b).

The Research Method

The data for this research were gathered during the Fall, Winter, and Spring quarters of the 1987-88 academic year. The COMP exam was administered to 1828 seniors, and the Academic Profile was administered to 1173 seniors. Assignment to a testing group was based on two criteria: first, all students who had taken the COMP exam as freshmen were assigned to the COMP testing group as seniors; second, students who were not tested as freshmen were randomly assigned to either the COMP or the Academic Profile testing groups. Also, during the Winter quarter 35 seniors agreed to take both exams. Since these students were volunteers, they were compensated for their participation.

Students' answer sheets were returned to the test developers who scored the exams and returned score and subscore reports to UTK. In addition, ACT returned to

UTK a data tape that contained students' responses to each of the items on the COMP exam. ETS did not provide item response data for the Academic Profile.

In addition to completing the exams, students were asked a series of questions about their educational experiences and their perceptions of the two exams. Answers to these questions were combined with data from UTK's administrative records to provide profiles of students' entering ability levels and patterns of coursework.

UTK administrative records were used to gather data on students' background characteristics and ability levels as measured by the ACT Assessment examinations, which consist of four tests: English Usage, Mathematics Usage, Natural Science Reading, and Social Studies Reading. A Composite score also is reported. The alpha reliability estimate for the Composite score is .85 and estimates for the four tests range from .73 to .77 (American College Testing Program, 1973).

Questionnaires administered to students after they had taken either the COMP exam or the Academic Profile were used to gather information about their perceptions of the exams and their patterns of coursetaking. Three questions about student perceptions of the tests queried students' opinions regarding the quality of the exams as measures of general education knowledge and skills, the quality of the exams as interesting experiences, and the students' levels of motivation to do well on the exams.

Questions about coursework differed for the two testing groups. Students in the COMP testing group answered questions developed at UTK. This group was asked to identify, from a list, courses they had taken in the areas of History, Humanities, Mathematics, Natural Science, and Social Science while attending UTK. Student answers consisted of yes-no responses.

Students in the Academic Profile testing group answered coursework questions developed by ETS. These questions covered the areas of Business and Commerce, Humanities, Science and Mathematics, and Social Science (Educational Testing Service, 1987). Students responded to these questions by indicating whether they had taken no courses, 1-2 courses, 3-6 courses, 7-9 courses, or 10 or more courses in a discipline.

While there are potential problems with reliance on self-report data, previous research on self-reports of coursetaking have revealed that coursework means for the ten undergraduate colleges at UTK are reasonably consistent with the coursework requirements of those colleges (Pike & Phillippi, 1988).

In order to identify patterns of coursetaking, principal components analyses were performed on the two sets of coursework variables. The results of these analy-

ses suggest that three coursework patterns are present for each testing group. These coursework patterns, along with the courses comprising each pattern, and alpha reliability estimates are presented in Table 1.

Insert Table 1 about here

An examination of the data presented in Table 1 suggests that the three coursework patterns identified for the COMP testing group represent coursetaking in calculus and the physical sciences, business mathematics and the social sciences, and the biological sciences and humanities. Alpha reliability coefficients for the scales representing these coursework patterns are .95, .90, and .77 respectively. The three coursework patterns for the Academic Profile testing group represent coursetaking in business and commerce, mathematics and the physical sciences, and humanities and the social sciences. Alpha reliability estimates for these scales are .93, .77, and .70 respectively.

Results

Sample Characteristics

An examination of the sample characteristics of the COMP exam and the Academic Profile testing groups reveals that the two groups are quite similar. Approximately 53% of the COMP testing group and 52% of the Academic Profile testing group are males; slightly more than 92% of the COMP testing group and 93% of the Academic Profile testing group are white; less than 4% of the COMP testing group and 3% of the Academic Profile testing group are black; and the corresponding percentages for Asian students are 3% and 4% respectively. Chi-square tests confirm that the two groups are not significantly different in terms of gender or race.

An examination of the background characteristics of the 35 students who took both the COMP exam and the Academic Profile reveals that 57% of these students are males and 43% are females, almost 92% are white, 5% are black, and the remaining 3% are Asian.

Some significant differences in the entering ability levels of these students are present, however. Group means, along with analysis of variance results, are presented in Table 2. An examination of the data reveals that students in the COMP testing group scored significantly higher on three of the four ACT Assessment examinations than did the Academic Profile group, and these differences are reflected in higher ACT Composite scores for that group. Differences in ACT scores are most

noticeable for the English Usage test (21.33 versus 20.65) and ACT Composite scores (22.09 versus 21.35). While the observed differences are statistically significant, the proportion of variance explained by the differences is trivial, suggesting that significant ANOVA results are due largely to sample size.

Insert Table 2 about here

The students who took both the COMP exam and the Academic Profile have a higher mean ACT Composite score (23.46) than the other two testing groups, a fact primarily due to a higher mean on the Mathematics Usage test (24.58). Although the mean high school grade point average for students taking both tests (3..7) is slightly higher than the means for the COMP exam and the Academic Profile testing groups, mean college grade point averages are the same for all three groups (2.87).

Patterns of coursetaking, as measured by the amount of coursework in an area, also are quite similar for the three groups. Means for both the COMP testing group and the Academic Profile testing group indicate that the average student takes between two and three courses per area. Although the coursework patterns of students who took both exams are similar to the other two groups, the amount of coursework in mathematics and the physical sciences is slightly higher for this group.

Student Performance

Table 3 presents all three groups of students' test scores and subscores on both the COMP exam and the Academic Profile along with the percentage of correct answers for these students, national mean percentage correct scores (norms), and ratios expressing the scores of UTK students in terms of national norms. Standard deviations also are included as indicators of score dispersion.

Insert Table 3 about here

At first glance, the data in Table 3 suggest that UTK students perform better on the COMP exam than on the Academic Profile since the mean total score for the COMP testing group is 187.62 (78% correct) as compared to a mean total score of 86.72 (60% correct) for the Academic Profile testing group. For the group that took both exams, the mean total score on the COMP exam is 189.97 (79% correct) and that

of the Academic Profile is 97.01 (67% correct). As the data in Table 3 indicate, similar differences are present for test subscores.

National averages for the two tests clearly show, however, that the COMP exam is a much less difficult test than the Academic Profile. The national average for total scores on the COMP exam is 185.2 (77% correct), while the estimated national average for total score on the Academic Profile is 72.0 (50% correct) (American College Testing Program, 1988a; Dick Burns, personal communication, January 9, 1989). A similar pattern is present for subscores on the exams.

When mean percentage correct scores for UTK students are expressed as ratios of national mean percentage correct scores, it becomes obvious that student performance on the more difficult Academic Profile is superior to student performance on the COMP exam. It would seem that for UTK students, more difficult exams tend to produce higher levels of motivation to do well on the test. For total scores, the ratio of UTK to national percentage correct scores is 1.01 of the COMP testing group and 1.20 for the Academic Profile testing group. Differences are even more dramatic for the students taking both exams. In this group, the ratio of UTK to national percentage correct scores is 1.03 for the COMP exam and 1.34 for the Academic Profile.

The data in Table 3 also suggest that there is greater variability in scores on the Academic Profile than on the COMP exam. For the Academic Profile testing group, the standard deviation for total scores is 21.69, while the standard deviation is 15.10 for the COMP testing group. Even greater differences are present for students taking both exams. For this group the standard deviation for total scores is 11.43 for the COMP exam and 19.13 for the Academic Profile.

The Substantive Component of Construct Validity

As previously noted, the first step in evaluating the content coverage of the COMP exam and the Academic Profile involved using a panel of faculty "experts" to compare the content of the tests to UTK's goals for general education. Results of the faculty members' evaluations are presented in Table 4.

Insert Table 4 about here

As the data in Table 4 indicate, evaluators believe that the Academic Profile is superior to the COMP exam in its coverage of basic skills (50% versus 36%). Within the basic skills domain, both tests are viewed as providing complete coverage

of reading and problem-solving skills, but the Academic Profile is seen as superior to the COMP exam in its coverage of basic skills related to composition and computation. Neither test covers foreign language or computer skills.

Coverage of the knowledge and processes domain is approximately the same for both tests (29% versus 25%) in that neither provides complete coverage of any of the knowledge components. Although the COMP exam is superior to the Academic Profile in its coverage of aesthetics, technology, and economics, the Academic Profile is superior to the COMP exam in its coverage of science in life and the social sciences. Neither test covers the areas of Western history or foreign culture.

Coverage of the attitudes and perceptions domain is poor for both tests (20% versus 10%). While both cover the area of political and social dynamics equally well, the COMP exam is superior to the Academic Profile in its coverage of values. Neither test covers personal wholeness, life-long learning, or experience in learning.

Overall, the COMP exam and the Academic Profile are equal in their coverage of UTK's general education domains (29% versus 30%). More important, though, content coverage, while equal, is relatively poor; many areas are poorly represented or are not covered at all. Despite these limitations, the faculty evaluators believe that either test can be used to assess general education outcomes as long as users keep the limitations firmly in mind.

In order to gather data about students' perceptions of the COMP exam and the Academic Profile as measures of general education knowledge and skills, students were asked to rate the content coverage of the exams as excellent, good, satisfactory, fair, or poor. Results indicate that less than half (48%) of the students taking the COMP exam alone rated that test as satisfactory or better and slightly less than 18% rated the exam as good or excellent. Of the students who gave the COMP less than a satisfactory rating, slightly more than half (27% of the total) gave the test a poor rating.

Students had slightly more negative perceptions of the Academic Profile. Approximately 44% of the students gave this test a rating of satisfactory or better and 16% gave it a rating of good or excellent. Half of the students giving the Academic Profile a less than satisfactory rating (28% of the total) rated the test as a poor measure of general education knowledge and skills.

Students who took both exams were more critical of the COMP exam and more favorably disposed toward the Academic Profile. Nearly 30% of these students rated

the COMP exam as excellent, good, or satisfactory, while 47% gave the Academic Profile a favorable rating.

In summary, both faculty and student evaluators are critical of the content coverage of both the COMP exam and the Academic Profile as measures of general education outcomes at UTK. Even though students' evaluations were more favorable than the evaluations of the faculty panel, a majority of the students still rated the two tests as fair or poor measures of general education knowledge and skills.

The Structural Component of Construct Validity

Evaluating the norm-referenced scoring procedures of the COMP exam and the Academic Profile is an important part of assessing the structural component of construct validity because of the THEC's reliance on norm-referenced scores in allocating performance-funding dollars. In order to determine if the use of norm group comparisons is valid, the comparison groups provided by the test developers were examined.

An examination of the comparison group for the COMP exam reveals that ACT compares the performance of UTK seniors to the performance of all other seniors taking the test. While the comparison group does include some major research institutions (e.g., Louisiana State University, Pennsylvania State University, University of Iowa, University of Oklahoma, etc.), this group is heavily weighted toward small liberal arts colleges (e.g., Cedarville College, Marion College, Our Lady of the Lake University, William Jewel College, etc.) and regional state universities (e.g., Eastern Illinois University, Northeast Missouri State University, Southeastern Oklahoma State University, etc.). This is not to say that the general education programs at these institutions are not comparable to those at UTK, but that no information about the comparability of these programs is provided by ACT.

Moreover, ACT does not provide any information about sampling procedures at comparison institutions. For example, James Madison University (JMU) is included in the comparison group and the sample at that institution consists of approximately 200 seniors. The mean ACT equivalent score for the JMU sample is approximately 25, as compared to an entering ACT score of 22 at UTK (T. Dary Erwin, personal communication, January 18, 1989). This three-point difference is extremely significant because a difference of one point in ACT scores translates into almost a ten percentile point difference in expected COMP scores (Steele, 1988).

An examination of the Academic Profile norm groups reveals that ETS has provided a comparison group consisting of all institutions using the exam, as well as

comparison groups consisting of institutions with the same Carnegie classifications. In addition, ETS allows institutions to select their own comparison groups. While ETS does provide greater flexibility in the selection of reference groups, it does not provide information about the general education programs at those institutions, the ability levels of students at those institutions, or sampling procedures at comparable institutions. In the case of UTK, the comparison group of institutions with the same Carnegie classification includes Washington State University, which used the Academic Profile as part of a statewide research project. The student sample at Washington State consisted of 200 volunteers who were paid for their participation (Thorndike, Gill, Gillmore, & Hunter, 1988). If the sample of students at UTK who took both exams is any indication, paid volunteers will have significantly higher scores than the general student population at an institution. Thus, not knowing about the characteristics of the samples included in the comparison group can result in invalid comparisons between institutions.

A second problem with the norm-referenced comparisons used by the THEC is the use of percentile ranks to compare programs to the norm group and to award funds. This problem is created because the COMP exam is a relatively easy test. The low difficulty levels for the COMP exam restrict the range of students' scores, producing relatively low levels of score variance. Low difficulty levels, coupled with low levels of score variance and the fact that each item on the COMP exam is worth four points, creates a situation in which small changes in student performance can have an enormous effect on percentile ranks. For example, a change in two responses on one of the items produces a change in the total score of four points (out of a possible 240 points). This four point score change is less than a 2% change in the possible score, but it translates into approximately a ten percentile point gain or loss for scores between the 30th and 70th percentiles. A score decline of approximately five points (which occurred between the Fall and Winter quarters at UTK), translates into almost a 15 percentile point decline from the 60th to the 47th percentiles.

A second important element in assessing the structural component of construct validity is the reliability/generalizability of scores on the COMP exam and the Academic Profile. Questions of reliability (internal consistency) are important because both the COMP exam and the Academic Profile rely on an additive scoring model and because unreliability is a source of irrelevant score variance.

In order to evaluate the reliability of the exams, alpha reliability coefficients were calculated for the total score and subscores on the COMP exam. Because

item responses were not available for the Academic Profile, alpha reliability estimates could not be calculated. However, by using score means and standard deviations (Gulliksen, 1950), it was possible to calculate KR-20 reliability estimates for this test. According to Gulliksen, the method used to estimate the reliability of Academic Profile scores is relatively conservative, especially if there is variation in item difficulty levels. Reliability coefficients and standard errors of measurement for the total scores and subscores of the COMP exam and the Academic Profile are presented in Table 5.

Insert Table 5 about here

An examination of these data indicates that the total scores and subscores of the Academic Profile are more reliable than the total scores and subscores of the COMP exam. Specifically, the KR-20 reliability estimate for total scores on the Academic Profile is .93, as compared to the alpha reliability coefficient of .76 for the total score on the COMP exam; KR-20 estimates for the Academic Profile subscales range from .79 to .84, while reliability estimates for the COMP exam range from 44 to .60. As would be expected, standard errors of measurement are greater for the COMP exam than for the Academic Profile. The standard error for COMP total scores is 7.40, compared to a standard error of 5.74 for total scores on the Academic Profile; standard errors of measurement for COMP subscales range from 3.92 to 4.88, and standard errors for the Academic Profile subscales range from 2.76 to 3.25.

It is worth noting that KR-20 reliability estimates for the total scores and subscores on the Academic Profile are quite close to the alpha reliability estimates reported by ETS. According to Gulliksen (1950), this correspondence between local KR-20 reliability estimates and national reliability estimates indicates that the difficulty levels of the questions on the Academic Profile are highly homogeneous for the UTK sample.

When assessing the effectiveness of academic programs for the THEC, institutional means, not individual scores, are the unit of analysis, and the generalizability of scores over items and over subjects is a paramount concern. At UTK, where different colleges have different general education requirements, it is important that scores be generalizable to colleges as well as to the institution as a whole. In the present research, generalizability was assessed, using procedures suggested by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and adapted by Kane, Gillmore, and Crooks (1976). The variance components used to calculate

generalizability coefficients were derived from a database of COMP scores for 30 colleges, each with samples of 300 students (Pike & Phillippi, 1989). While this database does not duplicate national data on the COMP exam exactly, individual and institutional means and standard deviations are reasonably close to those reported by ACT. Generalizability coefficients for UTK and its colleges were calculated, using the sample sizes in the present research. Because ETS does not provide item-level data, generalizability coefficients could be calculated for only the COMP exam and its subscores. The results of these analyses are presented in Table 6.

Insert Table 6 about here

An examination of the data in Table 6 reveals that it is feasible to generalize over both students and items to obtain a university mean for the total score on the COMP exam ($Ep^2_{(S,I)}$=.82). With the exception of the Using Science and Technology subscale ($Ep^2_{(S,I)}$=.77), it is not possible to generalize with confidence over students and items to obtain university means on the remaining COMP subscales. Generalizability coefficients for institutional means on the COMP subscales range from .39 for Solving Problems to .67 for Functioning within Social Institutions.

The data in Table 6 also indicate that the generalizability of total scores and subscores is influenced by sample size. Generalizability coefficients for colleges with more than 200 students in the sample (Business, Engineering, and Liberal Arts) are greater than .80, and generalizability coefficients for colleges with less than 50 students in the sample (Agriculture, Nursing, and Social Work) are less than .70. Increasing the sample size beyond 200 students has little practical effect on the generalizability of total scores since the maximum possible generalizability coefficient is .83 for an infinitely large sample.

The final column in Table 6 contains the 95% confidence intervals for institutional means, which were calculated by using procedures suggested by Cronbach, et al. (1972). As these data indicate, the confidence intervals for COMP total score means are within acceptable boundaries (±5.56). Using the sample sizes of 50 ($Ep^2_{(S,I)}$=.70) and 200 ($Ep^2_{(S,I)}$=.80) identified previously, the 95% confidence intervals are ±6.24 and ±5.72 respectively. With an infinitely large sample the 95% confidence interval for the COMP total score mean is ±5.54.

Examining the relationships among the subscales of an achievement test can provide important information about whether a test actually measures the outcomes it purports to measure. At issue in the present investigation is whether the subscales

of the COMP exam and those of the Academic Profile actually measure distinct, although possibly related, dimensions of general education.

The first step in evaluating the relationships among the subscales of the two tests involved calculating correlations among the subscales of each test independently. Because both the COMP exam and the Academic Profile rely on a matrix approach in their design (i.e., the same questions are used to measure content and process/skill areas), correlations were calculated for these areas separately.

Correlations among the subscales of each of these exams are presented in Table 7. Coefficients above the diagonals are Pearson product-moment correlations, and coefficients on the diagonals of each matrix are reliability estimates. The coefficients below the diagonals are correlations that have been corrected for attenuation (scale unreliability), using procedures suggested by Gulliksen (1950).

Insert Table 7 about here

An examination of the simple correlations in Table 7 indicates that the subscales of the COMP exam are less highly intercorrelated with each other than are the subscales of the Academic Profile. Correlations range from .45 to .55 for the COMP content subscales and from .48 to .49 for process subscales. In contrast, correlations for the subscales of the Academic Profile range from .76 to .83 for content areas and from .64 to .81 for skill areas.

One reason for lower intercorrelations among COMP subscales is that the COMP exam is a less reliable test than the Academic Profile, as indicated by the coefficients on the diagonal. When correlations are corrected for attenuation, the subscales of the two tests are almost perfectly correlated: disattenuated correlations range from .88 to .96 for the COMP content subscales and from .92 to 1.02 for the process subscales; those for the Academic Profile range from .92 to 1.00 for the content subscales and from .72 to 1.03 for the skill subscales. The one exception to this pattern is the Mathematics subscale on the Academic Profile, whose disattenuated correlations range from .72 to .81. The fact that some of the correlations are greater than unity suggests that items within one subscale are more highly correlated with items and subscores for another subscale than with items and subscores for their own subscale.

While an examination of the correlations among subscales suggests that there is a single dimension underlying the COMP exam and the Academic Profile, it is possible that other dimensions are present. In order to determine if unique aspects of

general education are being measured by the two tests, principal components analyses were performed. The results are presented in the four subtables of Table 8.

Insert Table 8 about here

The data in Table 8 provide strong support for the unidimensional structure of both the COMP exam and the Academic Profile. In no case does a meaningful second principal component emerge. For example, the analysis of COMP content subscores identified only one principal component with an eigenvalue greater than 1.0, and this component is able to explain 66% of the total variance. Results of analyses of the COMP process subscores, as well as the content and skill subscores of the Academic Profile, also indicate that only one principal component should be extracted. The proportion of total variance explained by the components is .66, .86, and .77 respectively. The finding that subscores are unidimensional is further supported by the fact that all subscales have significant positive pattern loadings on the first principal component.

Given the fact that both exams are unidimensional, the question arises as to whether there is any correspondence between the scores of the two tests. In order to answer this question, the scores of students taking both tests were intercorrelated. Table 9 presents the correlations between the subscales of the two tests.

Insert Table 9 about here

An examination of the upper left-hand portion of Table 9 reveals that content subscores on the two tests are positively correlated. Interestingly, the highest correlations are not for logical counterparts. For example, the Humanities subscale is more highly correlated with Using Science and Technology (.54) than with Using the Arts (.49). Similarly, the Natural Science subscale is more highly correlated with Using the Arts (.51) than with Using Science and Technology (.34). Based on these results, it seems safe to conclude that there is not a one-to-one correspondence between the content areas of the two tests.

Establishing a one-to-one correspondence between process/skill subscales is more difficult because the COMP exam and the Academic Profile differ in what they attempt to measure. It is worth noting that the Academic Profile Mathematics subscale is most highly correlated with Communicating (.57)on the COMP exam, the

subscale that contains mathematics questions. Two subscales that would be expected to be highly correlated are Critical Thinking and Solving Problems. However, the correlation between these two subscales is lower (.42) than that between Critical Thinking and Communicating (.52) or between Solving Problems and Writing (.46). Again, there does not seem to be a one-to-one correspondence between process/skill areas on the two tests.

The absence of a one-to-one correspondence between the subscales of the two exams clearly indicates that these subscales are not interchangeable. However, this does not mean that the two exams, in general, do not measure the same outcomes. In fact, the correlation between total scores on the two exams is .64. The correlations between the subscales of the two tests do suggest that the COMP exam and the Academic profile use slightly different approaches to measuring the same outcome. Indeed, the presence of a single dimension underlying both tests would help explain the significant correlations between natural sciences and the arts and between critical thinking and communicating.

In summary, the dominant finding concerning the dimensionality of the COMP exam and the Academic Profile is that both tests are unidimensional measures. Moreover, the significant correlations between scores on the tests suggest that the same dimension is being measured by both tests. While a precise identification of the outcome being measured by these tests must await research on the external component of construct validity, the results obtained thus far suggest that the outcome being measured is very similar to what Spearman (1904) terms "general intelligence."

The External Component of Construct Validity

The sensitivity of an achievement test to students' educational experiences is an important element in judging the appropriateness of that test as an assessment instrument. Questions related to the educational sensitivity of a test are central to demonstrating the construct validity of an outcomes measure. In order to evaluate the sensitivity of the COMP exam and the Academic Profile to educational experiences, scores on the two tests were correlated with measures of coursework, entering academic ability, and motivation when taking the test. These correlations are presented in Table 10.

Insert Table 10 about here

An examination of the correlations presented in Table 10 reveals some surprising relationships. For the COMP exam, calculus and physical science coursework is positively correlated with all scores. While it is understandable that this coursework pattern would be positively correlated with Using Science and Technology (.299) and Communicating (.253), it is surprising that it is also positively correlated with Functioning within Social Institutions (.126) and Using the Arts (.080). Furthermore, coursework in business mathematics and the social sciences is negatively correlated with all COMP scores, including Functioning within Social Institutions (-.046), and coursework in biology and the humanities is negatively correlated with Communicating (-.078) and positively correlated with Solving Problems (.088). The latter coursework pattern also is positively correlated with Using the Arts (.055) although the relationship is not significant.

While the influence of coursework on COMP scores is unclear, the relationships between ability and motivation measures and COMP scores are all positive and significant. For example, correlations between students' ACT scores and their total scores on the COMP exam range from .351 to .485, and the range of correlations between ACT scores and subscores is from .185 (ACT Mathematics and Solving Problems) to .492 (ACT Social Studies and Clarifying Values). Next to ability, motivation is most strongly and consistently associated with COMP scores. This association is strongest for total scores (.237) and ranges from .071 to .198 for the subscores.

For the Academic profile, business and commerce coursework and humanities and social science coursework are negatively related to all scores. In contrast, mathematics and physical science coursework is positively correlated with total scores (.135) and with the Natural Science (.270), Critical Thinking (.119), and Mathematics (.367) subscales. Mathematics and physical science coursework also is positively related to Humanities and Social Science subscores (.032 and .074 respectively) although these relationships are not significant.

As with the COMP exam, the strongest and most consistent relationships exist between test scores and ability and motivation measures. ACT scores and motivation are positively correlated with total scores on the Academic Profile (.456 to .559 and .373) and with all subscores on this exam (.281 to .590 and .259 to .381). These results suggest that both the COMP exam and the Academic Profile are first, measures of entering academic ability and second, measures of motivation to do well on the tests.

Because ability and motivation also may influence coursework, an attempt was made to isolate the unique effects of coursework on outcomes after controlling for

ability and motivation. In isolating the effects of coursework on outcomes, a causal model was specified and tested using LISREL (Joreskog & Sorbom, 1986). This model specified that ability and motivation measures were intercorrelated and influenced both coursework and outcomes. In turn, coursework variables were assumed to influence outcomes. Finally, residuals for the coursework variables were assumed to be intercorrelated, as were residuals for the subscores of the two tests. Separate analyses were conducted for total scores and for content and process/skill subscores. While the outcomes measures represented by this model may be a more accurate gauge of student learning than simple correlations, it should be remembered that the THEC awards funds on the basis of raw score improvement. Maximum likelihood estimates representing the effects of coursework, ability, and motivation measures on outcomes are presented in Table 11.

Insert Table 11 about here

The data in this table indicates that, at least for the COMP exam, controlling for the effects of ability and motivation clarifies the relationship between coursework and test scores. Taking courses in calculus and the physical sciences has a significant positive effect on Using Science and Technology scores (.123) and Communicating scores (.126); taking courses in business mathematics and the social sciences has a significant positive effect on Functioning within Social Institutions (.076); and taking courses in biology and the humanities has a positive effect on Using the Arts scores (.073) and Solving Problems scores (.095). Although the three coursework variables are positively related to total scores on the COMP exam, none of these effects is statistically significant. This finding is extremely significant because it suggests that actions to improve performance on COMP subscales will not be rewarded by improvements in overall performance.

Both ability and motivation significantly influence performance on the COMP exam. All four ACT Assessment scores are positively related to COMP total scores (.084 to .266) and subscores, and motivation to do well on the exam also influences total scores (.198) and all of the subscores (.087 to .219).

Untangling the relationship between coursework and scores on the Academic Profile is more difficult. Even after controlling for the effects of ability and motivation, business and commerce coursework, along with humanities and social science coursework, are negatively related to all scores, and mathematics and physi-

cal science coursework is negatively related to total scores (-.036) and all subscores except Natural Science (.075) and Mathematics (.143).

The effects of ability on Academic Profile scores ranges from .103 for ACT Mathematics to .260 for ACT English, and three of the four effects are greater for the Academic Profile than for the COMP exam. Motivation to do well on the exam has a significant maximum likelihood coefficient of .252 for Academic Profile total scores, as compared to a coefficient of .198 for COMP exam total scores. Thus, the effects of ability and motivation on Academic Profile scores are slightly stronger than similar effects on COMP scores.

Discussion

Before addressing the implications of these findings, a discussion of the limitations of the present research is necessary. The most basic limitation of this study concerns its purpose: to define a set of standards and a methodology for evaluating achievement tests as assessment instruments and to evaluate the utility of this method using data about two tests drawn from one institution.

This research was not intended to provide a definitive statement about the superiority of the COMP exam or the Academic Profile. Two factors limit the generalizability of these findings. First, this research was conducted at only one institution, and this institution has unique student characteristics, general education goals, and testing procedures. Thus, attempts to generalize the results of this research beyond the assessment program at UTK should not be made.

The second factor limiting the generalizability of the present research is the timeliness of the data. Since these data were collected, a new version of the COMP exam has been made available, and the Academic Profile has undergone substantial revision. The new Academic Profile has lower intercorrelations among its subscales, and ETS will be providing a form of criterion-referenced scoring, in addition to norm-referenced scores, to assist institutions in evaluating their general education programs (Dick Burns, personal communication, January 9, 1989).

A second limitation of this research is the unavailability of comparable data for both exams. Because ETS does not make item-level data available, some of the analyses could be conducted for the COMP exam, but not for the Academic Profile. The most obvious result is that the reliability and generalizability of Academic Profile scores could not be assessed directly. In addition, the lack of item-level data prevented any comparison of the item scoring models used for the two tests.

The lack of comparable coursework measures also handicapped the present research. At this point, it is not clear if the COMP exam is more sensitive than the Academic Profile to educational effects, or if the observed differences can be attributed to using different types of coursework measures.

While this paper does not endorse either the COMP exam or the Academic Profile, the present research clearly shows that the validation methodology described in this paper is workable, and that it can yield useful information about whether test content accurately represents the outcomes considered important by an institution. Furthermore, the standards for construct validity described in this paper can be used to evaluate the meaning of test scores and to judge their sensitivity to educational experiences.

In addition to serving as a criterion for selecting assessment instruments, the validation methodology presented in this paper can be used to assist in the design and functioning of an assessment program. For example, evaluating the substantive component of construct validity by examining the content representativeness of a test requires that faculty members familiarize themselves with the goals of an education program. This reexamination of educational goals is at least as valuable as testing students to determine what they have learned (Pike & Banta, 1987).

Evaluating the structural component of construct validity can also improve the practice of assessment by providing insight into score meaning and by giving assessment practitioners an idea of the confidence that can be placed in test scores. As was the case in this research, validation studies force assessment practitioners to examine their norm groups to determine if the institutions in those groups are comparable.

Finally, studies designed to evaluate the external component of construct validity can help guide improvement efforts while studies designed to evaluate the sensitivity of test scores can serve as models for further research intended to suggest actions to improve program quality. By designing an ongoing program to examine the relationships between students' educational experiences and their test scores, assessment practitioners can continually evaluate the validity of assessment instruments and monitor the effects of program changes.

Despite the caveats about the limitations of this study, the present research does provide useful information about the construct validity of the COMP exam and the Academic Profile as assessment instruments for use by UTK and the THEC. The results of this research are discussed in terms of their implications for construct underrepresentation and construct irrelevant score variance and are summarized in a

discussion of the social consequences of using either of these exams to evaluate program quality at UTK and throughout Tennessee.

From the perceptive of the Tennessee Higher Education Commission, questions about the construct validity of the COMP exam and the Academic Profile are a moot point because improved performance on the COMP exam is a definition of an effective general education program. However, at UTK and elsewhere, general education is more than a test score. Accordingly, this discussion of construct validity uses the general education goals at UTK as the basis for its evaluations and contrasts this construct definition with the uses of assessment data made by the THEC.

According to Humphreys (1986, p. 64), effective general education programs "prepare students for creative, rewarding lives and responsible participation as citizens of the nation and the world." Both the COMP exam and the Academic profile underrepresented those aspects of general education related to life after college - personal development and life-long learning - as well as those that seek to prepare students to be citizens of the world - goals related to a student's own cultural heritage and the cultures of others.

Both the COMP exam and the Academic Profile also fail to measure the multidimensional aspects of an effective general education program. If, as Warren (1988) observes, student learning is multidimensional, then it is essential that assessment instruments capture the important dimensions of this learning. Unfortunately, there is no evidence that either the COMP exam or the Academic Profile measure anything more than a single outcome of college.

Finally, the single outcomes dimension measured by the COMP exam and the Academic Profile does not seem to cover those general education outcomes that are sensitive to students' educational experiences. If an assessment instrument does not reflect students' college experiences, gathering assessment data is futile because the information gathered cannot be used to suggest strategies for improving the quality of general education programs.

In addition to inadequately representing the constructs underlying effective education programs, both the COMP exam and the Academic Profile contain numerous sources of construct irrelevant test variance. For example, problems with the validity of normative comparisons produce variations in national percentile ranks that are the result of including institutions with widely varying general education programs and/or unrepresentative student samples in the norm group. Moreover, the fact that national percentile ranks for the COMP exam are overly sensitive to small

score changes means that a substantial amount of the variation in percentile ranks is due to trivial score changes.

Error of measurement is also an important source of irrelevant score variance. While the Academic Profile is a highly reliable test, the error of measurement for total scores on this test is sufficient to spell the difference between an effective and an ineffective program based on the THEC's performance funding formulas. This criticism is even more appropriate for the COMP exam which has greater errors of measurement than does the Academic Profile. Observed errors of measurement are sufficient to explain the fifteen percentile point decline in UTK's COMP scores from the Fall to Winter quarters as chance variation unrelated to changes in program quality.

Perhaps the most important source of construct irrelevant score variance is the sensitivity of both the COMP exam and the Academic Profile to students' levels of ability and motivation. Not only do the effects of ability and motivation mask other relationships, such as the relationship between coursework and test scores, but programs may be judged as effective or ineffective solely on the basis of ability and/or motivation.

Given the limitations of both exams as assessment instruments, assuming that either test is the sine qua non of effective general education programs can have disastrous consequences. On one hand, if either of these tests is used as the sole basis for judging the effectiveness of a general education program, there is a real danger that important aspects of general education which are not measured will not be given the attention they deserve. What is important will come to be defined as what is easy to measure.

On the other hand, if aspects of general education not covered by the COMP exam or the Academic Profile are measured independently and results indicate that improvement actions should be taken, the University will be faced with a dilemma. Should the University spend some of its limited resources to improve general education in areas for which it will not be rewarded, or should the University spend its resources only in areas where rewards will be forthcoming?

The underrepresentation of students' educational experiences can also have undesirable consequences. The problem is most clearly seen in the case of the COMP exam. As previously noted, the THEC requires that institutions use the subscores on the COMP exam to guide corrective actions. The effectiveness of these corrective actions then are evaluated in terms of their impact on COMP total scores. The results of this research clearly show that program changes designed to improve

performance on a subscale of the COMP exam will not translate into improvements in total scores. While the THEC awards performance funding dollars simply for taking actions, the University is face with the choice of taking corrective actions that will gain funds for adopting the "proper" improvement measures but will not benefit the University in the long run by improving total scores, or it can take "improper" actions that will improve total scores but will not influence performance on the subscales. In the case of the Academic Profile, it is not clear that any actions can be taken that will improve performance either on the subscores or the total score.

The presence of construct irrelevant (error) variance in the COMP exam and the Academic Profile, coupled with the precision ascribed to the test scores by the THEC guidelines, can also have undesirable consequences. For example, the UTK mean on the COMP exam is 187.62 and the 95% confidence interval for this mean is $\pm$5.56 points, creating a range from 182.06 to 193.18. According to ACT (1988b), this translates into percentile ranks ranging from the 40th to somewhere between the 60th and 65th percentiles. This percentile range covers THEC funding scores of three to nine points. Since each score point is worth about $45,000 to UTK, an allocation ranging anywhere from $135,000 to $405,000 is possible by chance alone.

What is perhaps the most undesirable consequence of using either the COMP exam or the Academic Profile as the sole basis for evaluating the effectiveness of a general eduction program is the strong link between ability and test scores. If program effectiveness is defined in terms of scores on either of these tests, the quality of a program is the quality of the students attending the University. Basing judgments about program quality on measures of student quality is the very thing that the THEC was attempting to avoid when it adopted the performance funding standards (Levy, 1986). In a worst-case scenario, an institution would seek to improve its performance rating, not by improving its academic programs, but by becoming more selective and restricting admissions to the most able students.

One saving feature of the performance funding standard for general education is that it bases half of its funding award on value-added (gain) scores. Since entering ability is negatively related to gain scores (Banta, et al., 1987), the benefits of selective admissions on total scores tends to be partly offset by deficits in gain scores. Unfortunately, as Banta, et al. (1987) also have shown, actions designed to improve total scores (e.g., strong general education components and higher levels of student involvement) are negatively related to gain. Thus, the use of gain scores in conjunction with total scores means that the two measures will

offset each other, and virtually no action will have a lasting effect on program quality as defined by the performance funding guidelines.

In sum, the results of this research clearly indicate that much remains to be done at UTK and in Tennessee to define the outcomes that represent effective education programs, to identify or develop the instruments that accurately measure those outcomes, and to use assessment data in such a way as to reward effective education programs. Accomplishing these three objectives will help ensure that the THEC's goal of rewarding institutions based on what they do, not on who they recruit, is finally realized.

Conclusion

This paper makes the very basic argument that it is essential that the instruments used to assess student outcomes be valid measures of the constructs they are designed to evaluate. Validity in this case requires that test content accurately reflect the goals of an education program, that test structure reflect the structure of the outcomes being measured, and that test scores be sensitive to the educational experiences of students. As this research suggests, the continual validation of assessment instruments can be an expensive and time-consuming process. However, these are costs that must be borne if assessment is to realize its potential and serve as a catalyst for improving the quality of American higher education.

Notes

[1] Information on the pilot testing of the Academic Profile was obtained from Dick Burns, ETS College and University Programs, and is based on an analysis of approximately 1400 seniors who took the exam during Fall, 1987.

References

American College Testing Program. (1973). Assessing students on the way to college: Technical report for the ACT Assessment program. Iowa City: Author.

American College Testing Program. (1988a). College Outcome Measures Program: 1988-89. Iowa City, IA: Author.

American College Testing Program. (1988b). COMP senior reference group norms based on 40,625 seniors at 123 institutions [Unpublished data table]. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Banta, T. W. (1988a). Assessment as an instrument of state funding policy. In T. W. Banta (Ed.), Implementing outcomes assessment: Promise and perils (New directions for institutional research, No. 59, pp. 81-94). San Francisco: Jossey-Bass.

Banta, T. W. (1988b). Editor's notes. In T. W. Banta (Ed.), Implementing outcomes assessment: Promise and perils (New directions for institutional research, no. 59, pp. 1-4). San Francisco: Jossey-Bass.

Banta, T. W., & Fisher, H. S. (1987, March 4). Measuring how much students have learned entails much more than simply testing them. The Chronicle of Higher Education, pp. 44-45.

Banta, T. W., Lambert, E. W., Pike, G. R., Schmidhammer, J. L., & Schneider, J. A. (1987). Estimated student score gain on the ACT COMP exam: Valid tool for institutional assessment? Research in Higher Education, 27, 195-217.

Banta, T. W., & Schneider, J. A. (1988). Using faculty-developed exit examinations to evaluate academic programs. Journal of Higher Education, 59, 69-83.

Centra, J. (1988). Assessing general education. In C. Adelman (Ed.), Performance and judgement: Essays on principles and practices in the assessment of college student learning (OERI Publication No. OR 88-514, pp. 97-116). Washington, DC: U.S. Government Printing Office.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley.

Educational Testing Service. (1987). The Academic Profile: Information booklet. Princeton, NJ: Author.

El-Khawas, E. (1988). Campus trends, 1988. Washington, DC: American Council on Education.

Ewell, P. T., & Lisensky, R. (1988). Assessing institutional effectiveness: Redirecting the self-study process. Boulder, CO: National Center for Higher Education Management Systems.

Forrest, A., & Steele, J. M. (1982). Defining and measuring general education knowledge and skills. Iowa City, IA: American College Testing Program.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley.

Harris, J. (1976). Assessing outcomes in higher education. In C. Adelman (Ed.), Assessment in American higher education: Issues and contexts (OERI Publication No. OR 86-301). Washington, DC: U.S. Government Printing Office.

Humphreys, W. L. (1984). Interim report on general education. In T. W. Banta (Ed.), The NCHEMS/Kellogg student outcomes project at the University of Tennessee, Knoxville: Final report 1982-84 (pp. 51-60). Unpublished manuscript, University of Tennessee, Knoxville, Learning Research Center, Knoxville, TN.

Humphreys, W. L. (1986). Measuring achievement in general education. In T. W. Banta (Ed.), Performance funding in higher education: A critical analysis of Tennessee's experience (pp. 61-72). Boulder, CO: National Center for Higher Education Management Systems.

Joreskog, K. G., & Sorbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods (4th ed.). Mooresville, IN: Scientific Software.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 13, 171-183.

Levy, R. A. (1986). Development of performance funding criteria by the Tennessee Higher Education Commission: A chronology and evaluation. In T. W. Banta (Ed.), Performance funding in higher education: A critical analysis of Tennessee's experience (pp. 13-26). Boulder, CO: National Center for Higher Education Management Systems.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1987). Validity (ETS Research Report, No. RR-87-40). Princeton, NJ: Educational Testing Service.

Messick, S. (1988a). Meaning and values in test validation: The science and ethics of assessment (ETS Research Report, No. RR-88-47). Princeton, NJ: Educational Testing Service.

Messick, S. (1988b). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 167-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Millman, J. (1988). Designing a college assessment. In C. Adelman (Ed.), Performance and judgment: Essays on principles and practices in the assessment of college student learning (OERI Publication No. OR 88-514, pp. 9-38). Washington, DC: U.S. Government Printing Office.

National Governors' Association. (1988). Results in education: State-level college assessment initiatives - 1987-1988, results of a fifty-state survey. Washington, DC: Author.

Phillippi, R. H. (1989). A comparison of reliability and difficulty levels for three forms of the COMP exam. Unpublished manuscript, University of Tennessee, Knoxville, Assessment Resource Center, Knoxville, TN.

Pike, G. R. (1988a). Data on selected assessment instruments. In C. Adelman (Ed.), Performance and judgment: Essays on principles and practices in the assessment of college student learning (OERI Publication No. OR 88-514, pp. 313-325). Washington, DC: U.S. Government Printing Office.

Pike, G. R. (1988b, May). Students' background characteristics, educational experiences, and educational outcomes: A model for evaluating assessment instruments. Paper presented at the Conference on Developing Partnerships Between Community and Senior Colleges, Virginia Beach, VA.

Pike, G. R., & Banta, T. W. (1987). Assessing student educational outcomes: The process strengthens the product. VCCA Journal, 2(2), 24-35.

Pike, G. R., & Phillippi, R. H. (1988, November). Relationships between self-reported coursework and performance on the ACT-COMP exam: An analysis of the generalizability of the differential coursework methodology. Paper presented at the annual meeting of the Association for the Study of Higher Education, St. Louis, MO.

Pike, G. R., & Phillippi, R. H. (1989, May). Using generalizability theory in institutional research. Paper presented at the annual meeting of the Association for Institutional Research, Baltimore.

Spearman, C. (1904). "General intelligence," objectively determined and measured. American Journal of Psychology, 15, 201-292.

Steele, J. M. (1988). Report of results for the University of Tennessee, Knoxville, on use of the COMP Objective Test (Form VIII equated to Form III): September, 1988. Unpublished manuscript, American College Testing Program, Iowa City, IA.

Tennessee Higher Education Commission. (1987, July). Guidelines for performance funding. Unpublished manuscript, Nashville, TN.

Thorndike, R. M., Gill, J., Gillmore, G., & Hunter, S. (1988). Comparative evaluations of current standardized academic skills measures: Design of a state-wide study. Unpublished manuscript, Western Washington University, Bellingham, WA.

University of Tennessee Office of Management Services. (1987). A graphic view of the University of Tennessee (9th ed.). Knoxville, TN: Author.

Warren, J. (1988). Cognitive measures in assessing learning. In T. W. Banta (Ed.), Implementing outcomes assessment: Promise and perils (New directions for institutional research, No. 59, pp. 29-40). San Francisco: Jossey-Bass.

Table 1

Coursework Patterns, Courses, and Reliability Estimates for the COMP Exam and the Academic Profile Testing Groups

COMP TESTING GROUP		
Physical Sciences and Calculus (r_{xx}=.95)	Business Mathematics and Social Sciences (r_{xx}=.90)	Biological Sciences and Humanities (r_{xx}=.77)
Basic Engineering	Astronomy	Biology
Chemistry	College Algebra	Botany
Physics	Business Calculus	Microbiology
Freshman Calculus I	Math of Finance	Zoology
Freshman Calculus II	Literature	Art History
Freshman Calculus III	Speech and Theatre	Studio Art
Sophomore Calculus I	Economics	Dance
Sophomore Calculus II	Geography	Music History
Sophomore Calculus III	Political Science	Philosophy
	Psychology	Religious Studies
	Sociology	Anthropology
	Western Civilization	Child and Family Studies
		Social Work
		American History
		Cultural Studies

Table 1 continued

ACADEMIC PROFILE TESTING GROUP		
Business and Commerce (r_{xx}=.88)	Mathematics and Physical Sciences (r_{xx}=.77)	Humanities and Social Sciences (r_{xx}=.70)
Accounting	Chemistry	English
Economics	Computer Science	Fine Arts
Management	Engineering	Foreign Language
Marketing	Mathematics	Other Humanities
Other Business	Physics	History
		Psychology
		Sociology
		Other Social Sciences
		Biological Sciences
		Other Sciences
		Other Subject Areas

Table 2

COMP Exam and Academic Profile Testing Group Means and ANOVA Results for Selected Ability Measures

Ability Measures	GROUP MEANS		ANOVA RESULTS	
	COMP Exam	Academic Profile	F	Eta^2
ACT English	21.33	20.65	13.45***	.01
ACT Mathematics	22.10	21.56	4.79*	.00
ACT Natural Sciences	21.60	21.21	2.35	.00
ACT Social Studies	24.00	23.49	5.19*	.00
ACT Composite	22.09	21.35	14.37***	.01
High School GPA	3.16	3.11	5.95*	.00
College GPA	2.87	2.87	0.07	.00

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 3

Scores, Percent Correct Scores, UTK/Norm-Group Ratios, and Standard Devisations for Total Scores and Subscores on the COMP Exam and the Academic Profile

	COMP TESTING GROUP									
	Raw Score	Pct. Crrct.	Norm	UTK/ Norm	S_x	Raw Score	Pct. Crrct.	Norm	UTK/ Norm	S_x
Total	187.62	78	77	1.01	15.10	189.97	79	77	1.03	11.43
FSI	62.11	78	78	1.00	6.53	62.71	78	78	1.00	5.78
US	64.38	80	78	1.03	6.20	65.40	82	78	1.05	5.53
UA	61.07	76	75	1.01	5.77	61.83	77	75	1.03	4.94
COM	53.19	74	73	1.01	7.28	54.71	76	73	1.04	5 89
SP	76.32	80	78	1.03	6.48	76.83	80	78	1.03	5.51
CV	58.11	81	79	1.03	5.47	58.97	82	79	1.04	4.88
	ACADEMIC PROFILE TESTING GROUP									
	Raw Score	Pct. Crrct.	Norm	UTK/ Norm	S_x	Raw Score	Pct. Crrct.	Norm	UTK/ Norm	S_x
Total	86.72	60	50	1.20	21.69	97.01	67	50	1.34	19.13
HUM	29.08	61	50	1.22	7.95	33.55	70	50	1.40	6.98
SS	28.87	60	50	1.20	7.65	31.57	66	50	1.32	6.12
NS	28.77	60	50	1.20	7.75	32.02	67	50	1.34	7.78
READ	22.43	62	50	1.24	6.20	25.03	70	50	1.40	5.05
WRITE	23.27	65	50	1.30	6.17	26.00	72	50	1.44	5.19
CT	19.46	54	50	1.08	6.30	21.53	60	50	1.20	6.02
MATH	21.56	60	50	1.20	6.11	24.42	68	50	1.36	5.81

Table 4

Correspondence Between Test Content and UTK General Education Goals

General Education Goal	COMP Exam	Academic Profile
I. Basic Skills		
1. Verbal Communication		
A. English Composition	0%	50%
B. Spoken English	25%	0%
C. Reading Skills	100%	125%
2. Computational Skills	25%	75%
3. Foreign Language Skills	0%	0%
4. Computer Skills	0%	0%
5. Problem Solving	100%	100%
AVERAGE FOR BASIC SKILLS	36%	50%
II. Knowledge and Process		
1. Aesthetics	75%	50%
2. Science for Life	25%	50%
3. Technology	50%	25%
4. Western History	0%	0%
5. Foreign Culture	0%	0%
6. Economics	25%	0%
7. Social Sciences	25%	50%
AVERAGE FOR KNOWLEDGE AND PROCESS	29%	25%
III. Attitudes and Perceptions		
1. Values	50%	0%
2. Political Dynamics	50%	50%
3. Personal Wholeness	0%	0%
4. Life-long Learning	0%	0%
5. Experience in Learning	0%	0%
AVERAGE FOR ATTITUDES AND PERCEPTIONS	20%	10%
OVERALL AVERAGE	29%	30%

Table 5

Reliability Coefficients and Standard Errors of Measurement for Total Scores and Subscores on the COMP Exam and the Academic Profile

COMP TESTING GROUP		
Scale/Subscale	Reliability Coefficient	Standard Error
Total Score	.76	7.40
Functioning within Social Institutions	.54	4.43
Using Science and Technology	.60	3.92
Using the Arts	.45	4.28
Communicating	.55	4.88
Solving Problems	.51	4.54
Clarifying Values	.44	4.09

ACADEMIC PROFILE TESTING GROUP		
Scale/Subscale	Reliability Coefficient	Standard Error
Total Score	.93	5.74
Humanities	.84	3.18
Social Sciences	.82	3.25
Natural Sciences	.83	3.20
Reading	.80	2.77
Writing	.80	2.76
Critical Thinking	.79	2.89
Mathematics	.79	2.80

Table 6

Generalizability Coefficients for COMP Total Score and Subscores Given Sample Sizes for UTK and its Colleges

		GENERALIZABILITY COEFFICIENTS						
College	N	Total	FSI	US	UA	COM	SP	CV	C.I.
UTK (Total Sample)	1828	.82	.67	.77	.57	.61	.39	.56	5.56
Agriculture	27	.63	.51	.65	.38	.42	.26	.42	6.77
Archetecture	73	.74	.60	.72	.48	.52	.33	.50	6.03
Business	450	.81	.66	.76	.56	.59	.38	.55	5.62
Communications	127	.78	.63	.74	.52	.56	.35	.53	5.83
Education	137	.78	.63	.74	.52	.56	.36	.53	5.80
Engineering	285	.81	.65	.75	.55	.58	.38	.55	5.67
Human Ecology	124	.78	.63	.74	.52	.56	.35	.52	5.83
Liberal Arts	461	.81	.66	.76	.56	.59	.38	.55	5.62
Nursing	38	.68	.55	.68	.42	.46	.29	.46	6.44
Social Work	6	.35	.28	.32	.18	.20	.11	.23	11.22

Table 7

Correlations, Reliability Coefficients, and Disattenuated Correlations Among Subscales of the COMP Exam and the Academic Profile

	COMP EXAM					
Subscale	FSI	US	UA	COM	SP	CV
Functioning with Social Inst.	.54	.55	.45			
Using Science and Technology	.96	.60	.46			
Using the Arts	.92	.88	.45			
Communicating				.55	.49	.48
Solving Problems				.92	.51	.58
Clarifying Values				.98	1.02	.44

	ACADEMIC PROFILE						
Subscale	HUM	SS	NS	READ	WRITE	CT	MATH
Humanities	.84	.83	.76				
Social Sciences	1.00	.82	.80				
Natural Sciences	.92	.98	.83				
Reading				.80	.79	.81	.57
Writing				.99	.80	.75	.58
Critical Thinking				1.03	.95	.79	.64
Mathematics				.72	.73	.81	.79

Above Diagonal = Correlations

Diagonal = Reliability Coefficients

Below Diagonal = Disattenuated Correlations

Table 8

Results of the Principal Components Analyses for the Subscales of the COMP Exam and the Academic Profile

Principal Component	Eigenvalue	Explained Variance	Subscore	Pattern Loading
		COMP Content Subscores		
1	1.97	.66	FSI	0.83
2	0.58	.19	US	0.83
3	0.45	.15	UA	0.77
		COMP Process Subscores		
1	1.97	.66	COM	0.81
2	0.53	.18	SP	0.81
3	0.51	.17	CV	0.80
		Academic Profile Content Subscores		
1	2.59	.86	HUM	0.93
2	0.24	.08	SS	0.94
3	0.17	.06	NS	0.92

Table 8 continued

Academic Profile Skill Subscores				
Principal Component	Eigenvalue	Explained Variance	Subscore	Pattern Loading
1	3.08	.77	READ	0.91
2	0.50	.12	WRITE	0.89
3	0.25	.06	CT	0.92
4	0.18	.04	MATH	0.78

Table 9

Correlations Between the COMP and Academic Profile Subscales

ACADEMIC PROFILE	COMP EXAM					
	FSI	US	UA	COM	SP	CV
Humanities	.34	.54	.49	.51	.56	.36
Social Sciences	.29	.56	.53	.53	.52	.40
Natural Sciences	.15	.34	.51	.51	.27	.31
Reading	.26	.47	.43	.47	.40	.36
Writing	.27	.45	.43	.38	.46	.38
Critical Thinking	.26	.45	.54	.52	.42	.39
Mathematics	.15	.43	.50	.57	.39	.21

Table 10

Correlations Between COMP Exam and Academic Profile Soces and Subscores and Selected Coursework and Background Variables

	COMP EXAM						
Variables	Total	FSI	US	UA	COM	SP	CV
C/PS	.189***	.126***	.229***	.080*	.253***	.045	.128***
B/SS	-.124***	-.046	-.145***	-.090**	-.136***	-.045	-.102***
B/HUM	.013	-.020	-.004	.055	-.078*	.088**	.035
ACT/E	.421***	.367***	.283***	.336***	.364***	.377***	.296***
ACT/M	.351***	.277***	.291***	.250***	.393***	.185***	.225***
ACT/NS	.485***	.398***	.389***	.347***	.374***	.380***	.384***
ACT/SS	.445***	.353***	.415***	.273***	.305***	.368***	.492***
MOTIV.	.237***	.194***	.198***	.160***	.122***	.187***	.071*

	ACADEMIC PROFILE							
Variables	Total	HUM	SS	NS	READ	WRITE	CT	MATH
B/COM	-.153***	-.130**	-.122**	-.170***	-.172***	-.108*	-.176***	-.066
M/PS	.135**	.032	.074	.270***	-.021	-.006	.119**	.367***
H/SS	-.128**	-.047	-.117*	-.189***	-.032	-.032	-.107*	-.263**
ACT/E	.549***	.577***	.503***	.431***	.479***	.544***	.471***	.379***
ACT/M	.456***	.374***	.405***	.479***	.295***	.281***	.388***	.590***
ACT/NS	.545***	.492***	.515***	.495***	.430***	.414***	.551***	.462***
ACT/SS	.559***	.478***	.498***	.567***	.432***	.361***	.549***	.564***
MOTIV.	.373***	.335***	.330***	.364***	.322***	.309***	.381***	.259***

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 11

Maximum Likelihood Estimates for the Effects of Selected Coursework and Background Variables on COMP Exam and Academic Profile Scores and Subscores

	COMP EXAM						
Variables	Total	FSI	US	UA	COM	SP	CV
C/PS	.031	.047	.123**	-.030	.126**	-.022	.025
B/SS	.012	.076*	.023	-.026	.044	.020	-.001
B/HUM	.053	.013	.054	.073*	-.002	.095**	.056
ACT/E	.153***	.174***	.029	.166***	.142***	.136***	.077**
ACT/M	.084*	.045	.039	.093*	.173***	-.008	-.005
ACT/NS	.266***	.214***	.192***	.213***	.139***	.266***	.227***
ACT/SS	.151***	.111**	.222***	.016	.086*	.095*	.182***
MOTIV.	.198***	.163***	.174***	.130***	.219***	.087*	.157***

	ACADEMIC PROFILE							
Variables	Total	HUM	SS	NS	READ	WRITE	CT	MATH
B/COM	-.100***	-.071*	-.082*	-.117**	-.134***	-.055	-.124***	-.028
M/PS	-.036	-.079*	-.085*	.075*	-.144***	-.070	-.050	.143***
HUM/SS	-.060	-.002	-.061	-.084*	-.033	-.014	-.056	-.102**
ACT/E	.260***	.355***	.227***	.133**	.258***	.399***	.161***	.070
ACT/M	.103*	.050	.096*	.141**	.021	-.011	.039	.299***
ACT/NS	.198***	.175***	.224***	.138**	.164**	.163**	.272***	.074
ACT/SS	.222***	.166***	.180***	.272***	.191*	.068	.241***	.271***
MOTIV.	.252***	.226***	.225***	.246***	.224***	.219***	.271***	.147***

* $p < .05$; ** $p < .01$; *** $p < .001$

Figure 1

Messick's Facets of Validity

	Interpretation	Use
Evidence	Construct Valitity	Construct Validity & Relevance/Utility
Consequence	Construct Validity & Value Implications	Construct Validity Relevance/Utility Value Implications & Social Consequences